Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods

Year: 2022 | Citations: 148 | Authors: Evan Crothers, N. Japkowicz, H. Viktor

---

Abstract

Machine-generated text is increasingly difficult to distinguish from text authored by humans. Powerful open-source models are freely available, and user-friendly tools that democratize access to generative models are proliferating. ChatGPT, which was released shortly after the first edition of this survey, epitomizes these trends. The great potential of state-of-the-art natural language generation (NLG) systems is tempered by the multitude of avenues for abuse. Detection of machine-generated text is a key countermeasure for reducing the abuse of NLG models, and presents significant technical challenges and numerous open problems. We provide a survey that includes 1) an extensive analysis of threat models posed by contemporary NLG systems and 2) the most complete review of machine-generated text detection methods to date. This survey places machine-generated text within its cybersecurity and social context, and provides strong guidance for future work addressing the most critical threat models. While doing so, we highlight the importance that detection systems themselves demonstrate trustworthiness through fairness, robustness, and accountability.